

4 Simulace jako nástroj srovnávání metod práce s chybějícími hodnotami

V předchozích kapitolách jsme postupně definovali mechanismy chybějících hodnot a představili několik metod práce s chybějícími hodnotami. Následující tři kapitoly obsahují vlastní analýzy, kdy aplikujeme vybrané metody v rozličných situacích. Tyto situace jsou určeny primárně mechanismy chybějících hodnot. Hlavním cílem analytické části knihy však není pouze ukázat aplikaci těchto technik, ale pokusit se porovnat „úspěšnost“ vybraných metod práce s chybějícími hodnotami v různých situacích.

Analytickým nástrojem, pomocí kterého porovnávané metody práce s chybějícími hodnotami, je simulace. Simulace umožňují v rámci kontrolovaných podmínek imitovat reálné situace, se kterými je možné se běžně setkat při praktické analýze dat. V rámci přesně definovaných experimentálních podmínek tak simulace umožňují sledovat „úspěšnost“ jednotlivých metod práce s chybějícími hodnotami.

Tato kapitola definuje obecný rámec pro simulační studie realizované v dalších dvou kapitolách. V kapitole jsou nejdříve prezentovány motivace pro použití simulací jako analytického nástroje (podkapitola 4.1). Následně jsou představeny cíle analytické části knihy (podkapitola 4.2). Potom v krátkosti popisujeme jednotlivé technické úkony spojené s realizací simulačních studií (podkapitola 4.3) a statistický software, v kterém byly tyto kroky provedené (podkapitola 4.4). V závěru kapitoly jsou popsány nástroje pro vyhodnocování výsledků simulací (podkapitola 4.5).

4.1 Motivace pro používání simulací při srovnávání metod

Jelikož z podstaty věci chybějící hodnoty neznáme, tak nemůžeme vhodnost použití jednotlivých metod vyhodnocovat na reálných sociologických datech. Při vyhodnocování vhodnosti použití různých metod při rozličných statistických analýzách se proto používají simulovaná data (Little a Rubin 2002: 23). Zjednodušeně řečeno, po vytvoření kompletních simulovaných dat (tzn. dat bez chybějících hodnot) jsou z datové matice vymazané hodnoty a na takto redukováných datech se následně testují různé metody práce s chybějícími hodnotami. Vzhledem k potenciálnímu riziku, že v jediné takto vygenerované datové matici by v důsledku náhody mohly uspokojivé výsledky vykazat obecně nevhodné metody, se celý postup musí mnohokrát opakovat. Po dostatečném počtu opakování je na základě simulací možné dospět k validním závěrům.

Pro simulace s chybějícími hodnotami, které v analytické části realizujeme, je možné v metodologicko-statistické literatuře najít různé názvy. Někde bývají označené jako Monte Carlo simulace (např. Graham 2012: 229–239; Enders 2010: 52–53), jinde jako simulační studie (např. Collins et al. 2001: 339–347; Little a Rubin 2002: 23; Little 1988: 1201 anebo Fox 2008: 555) anebo jednoduše jako simulace (Schafer a Graham 2002). Všechny tyto „simulace“ však mají společné, že v rámci kontrolovaných podmínek zkoumají získané výběrové rozdělení relevantních statistik. Ačkoli výběrové

rozdělení statistiky vzhledem k jeho definici¹⁶⁵ představuje teoretický koncept, tak právě Monte Carlo simulace umožňují s využitím počítačů vytvořit aproximace výběrových rozdělení statistik (Paxton et al. 2001: 289). Navzdory odlišným názvům pro takového studie používáme v knize tři výše představené termíny jako synonyma.

Simulace všeobecně představují poměrně komplexní analytický nástroj. Komplexnost přitom závisí na množství variovaných experimentálních podmínek (tzn. podmínek, jejichž vliv v rámci definované výzkumné otázky zkoumáme). Platí přitom, že míra komplexnosti se zvyšuje s počtem variovaných podmínek (a přirozeně i s počtem variovaných hodnot v rámci jednotlivých podmínek). Komplexnost simulačních studií se projevuje též v množství na sebe logicky navazujících kroků, které musí být jednak teoreticky zdůvodněné, a zároveň také prakticky realizovatelné.

Je důležité podotknout, že Monte Carlo simulace je možné realizovat bez jakékoli souvislosti s věcnými fenomény (tzn. s hodnotami populačních parametrů, které nemusí mít žádné věcné opodstatnění¹⁶⁶). Jinými slovy, úlohu je možné nadefinovat výhradně matematicky¹⁶⁷, bez jakéhokoli vztahu ke společenským fenoménům. Vzhledem k tomu, že problematiku chybějících hodnot chceme představit společenskovední veřejnosti, jsme však zvolili věcně opodstatněné úlohy. Při první simulační studii pracujeme s proměnnými příjem osoby a hodnota IQ. Počítáním míry asociací mezi těmito proměnnými tak vlastně imitujeme zkoumání vztahu mezi příjmem osoby a vrozenou inteligencí (měřenou výsledkem standardizovaného testu). Druhá simulační studie imituje zkoumání determinantů politické znalosti, když jako vysvětlující proměnné tohoto konstruktů používá pohlaví, dosažené vzdělání a zájem o politiku.

4.2 Cíle analytické části

V analytické části knihy zkoumáme vhodnost použití šesti vybraných metod práce s chybějícími hodnotami při ignorování jakýchkoli informací o samotných chybějících hodnotách v datech a také teoretických poznatků o fungování těchto metod. Jinými slovy řečeno, zkoumáme následující hypotetickou situaci: „Co se může stát, pokud výzkumník bez zvažení potenciálních hrozeb arbitrárně zvolí některou z technik?“ Jednou z implikací takového definování „výzkumné otázky“ proto je, že všechny metody jsou aplikované i na situace (určené mechanismem chybějících hodnot), pro které nejsou (na základě v teoretické části knihy citované metodologické literatury) vůbec vhodné.

¹⁶⁵ Výběrové rozdělení statistiky je pravděpodobnostní rozdělení hodnot statistiky ve všech možných výběrech se stejným rozsahem z dané populace (Hendl 2006: 150).

¹⁶⁶ Příklady takovýchto simulačních studií, které se týkají problematiky chybějících hodnot, je možné nalézt například u Collinse a spoluautorů (2001), Foxe (2008: 554–555) a Littlea (1988).

¹⁶⁷ Mohli bychom například počítat korelační koeficient (spolu s dalšími charakteristikami) mezi proměnnými s dvourozměrným normálním rozdělením, kde by obě proměnné měly libovolně zvolené hodnoty průměru a rozptylů. Zároveň by mohly být arbitrárně zvoleny i hodnoty kovariance/korelace mezi těmito proměnnými.

K takovému nastavení analytické části nás vedou zjištění přehledových studií, které uvádějí, že problematika chybějících hodnot bývá ve vědeckých člancích ze společenskovední oblasti málo reportována. Konkrétně, autoři těchto článků většinou vůbec neuvádějí, jestli a případně do jaké míry se v datech, na kterých byly realizovány věcné analýzy, vyskytují chybějící hodnoty. Ještě v menší míře přitom bývají reportovány použité metody práce s chybějícími hodnotami. Například King a kolegové (King et al. 2001: 49) na základě obsahové analýzy článků ze tří nejprestižnějších politologických časopisů¹⁶⁸ konstatují, že jen v 19 % analyzovaných článků autoři explicitně popisují způsob nakládání s chybějícími hodnotami. Zároveň uvádějí, že v přibližně 94 % „kvantitativních“ článků s chybějícími hodnotami v datech jejich autoři pracovali jen s kompletními případy (tzn. že případy s chybějícími hodnotami alespoň u jedné z proměnných do analýzy vůbec nevstoupily). Dále Peugh a Enders (2004: 541–542) uvádějí, že v rámci 545 náhodně vybraných studií z oblasti výzkumu vzdělávání jen 42 % (229) prokazatelně obsahovalo chybějící hodnoty¹⁶⁹. Pozitivně je přitom možné hodnotit, že u těchto studií již v téměř 75 % případů autoři explicitně uvedli výskyt chybějících hodnot. V rámci uvedených 229 studií však jen 6 pracovalo s mnohonásobnými imputacemi anebo metodami založenými na maximální věrohodnosti (Peugh a Enders 2004: 542). Dramatičtější údaje reportují McKnight s kolegy (2007: 3), když v rámci 300 článků z oblasti psychologie přibližně 90 % prokazatelně obsahovalo chybějící hodnoty. Jen malý podíl však explicitně zohlednil problém výskytu chybějících hodnot, přičemž ještě menší podíl uvedl řešení tohoto problému¹⁷⁰.

Cílem analytické části knihy je pokusit se pomoci v dalších kapitolách dvou podrobně definovaných simulačních studií odpovědět na čtyři otázky, které jsou položeny v následujících odstavcích. Tyto odstavce vždy začínají formulací výzkumné otázky a za každým následuje věcné opodstatnění dané otázky. Připomínáme, že se stále pohybujeme v rámci hypoteticky nastavené situace, kdy se prostřednictvím aplikování zvolených metod práce s chybějícími hodnotami za rozličných okolností¹⁷¹, pokoušíme kvantifikovat potenciální nevýhody těchto technik.

Jaký podíl chybějících hodnot v datech je při jednotlivých mechanismech chybějících hodnot problematický? Kladení této otázky je relevantní, protože metodologická literatura jednoznačně nespecifikuje, jaký podíl případů s chybějícími hodnotami je při analýze problematický. Potenciální hrozba vychýlení odhadovaných parametrů totiž nezávisí jen na podílu chybějících hodnot, ale také na rozptylu zkoumaných proměnných, kovariancemi mezi proměnnými (McKnight et al. 2007: 139), mechanismu chy-

¹⁶⁸ Jedná se o tyto časopisy: *American Political Science Review*, *American Journal of Political Science* a *British Journal of Political Science*. Analyzované články byly publikovány v letech 1993 až 1997. Obsahová analýza reportování problému chybějících hodnot se přirozeně týkala jen článků, které obsahovaly kvantitativní analýzy z výběrových výzkumů (King et al. 2001: 49).

¹⁶⁹ Skutečný podíl je však z důvodu nemožnosti zjistit existenci chybějících hodnot v mnohých člancích vyšší (Peugh a Enders 2004: 541).

¹⁷⁰ Přesné podíly tyto autoři bohužel neuvádějí.

¹⁷¹ Přesněji se zajímáme o fungování uvedených metod ve všech situacích, které jsou určené kombinacemi variovaných experimentálních podmínek.

běžících hodnot (Schlomer et al. 2010: 2) a případně dalších faktorech. Proto v analytické části knihy při jednotlivých mechanismech chybějících hodnot variujeme podíl případů, u kterých hodnoty nebyly pozorované. Zároveň chceme ukázat, jakým způsobem se při použití zvolených metod práce s chybějícími hodnotami tato problematičnost realizuje (ve vztahu k odhadovaným parametrům a statistickému usuzování).

Za jakých okolností poskytují srovnávané metody práce s chybějícími hodnotami nevychýlené odhady populačních parametrů? Tato otázka přímo navazuje na předchozí otázku, přičemž však klade větší důraz na celkovou situační podmíněnost analýzy dat, které obsahují chybějící hodnoty. Konkrétně se tedy zajímáme o poměrně komplexní situace, které jsou definované kombinací čtyř podmínek: podílem chybějících hodnot, mechanismem chybějících hodnot, metodou práce s chybějícími hodnotami a samotným odhadovaným parametrem. Je totiž možné, že některé metody za určitých okolností (určených podílem chybějících hodnot a mechanismem chybějících hodnot) poskytují nevychýlené odhady určitých parametrů, přičemž při jiných parametrech už mohou vést k významným vychýlením. Takto položená otázka je podle nás opodstatněná, jelikož cílem různých analýz může být odhad různých parametrů. Pokud však vybraná metoda poskytuje za daných okolností nevychýlené odhady všech zkoumaných parametrů, jedná se (alespoň z hlediska vychýlení) o optimální metodu.

Umožňují dané techniky dospět ke správným úsudkům o populačních parametrech na základě výběrových dat? Při kladení této otázky soustředíme naši pozornost na směrodatné chyby a jejich potenciální vychýlení způsobené jednotlivými metodami práce s chybějícími hodnotami. Směrodatná chyba (anglicky *standard error*) představuje směrodatnou odchylku výběrového rozdělení statistiky (Everitt a Skrondal 2010: 409). Tato charakteristika, která zjednodušeně řečeno „měří nejistotu“ spojenou s usuzováním o populačních parametrech na základě výběrových statistik (Diez et al. 2012: 163), představuje klíčový koncept klasické frekvenční inferencí. Směrodatné chyby se totiž nachází „v jádru“ dvou hlavních nástrojů frekvenčního přístupu ke statistickému usuzování. Konkrétně tím máme na mysli intervaly spolehlivosti a testy významnosti (Everitt a Skrondal 2010: 174). Směrodatné chyby se vyskytují ve vzorcích na výpočet intervalů spolehlivosti (Hendl 2006: 172) a testovacích statistik (Hendl 2006: 177). Některé metody práce s chybějícími hodnotami sice mohou vést k nevychýleným odhadům parametrů, ale zároveň mohou produkovat vychýlené (většinou podhodnocené) hodnoty směrodatných chyb. Pokud jsou tedy hodnoty směrodatných chyb na základě výběrových dat vychýlené, tak u těchto metod nominální hladiny spolehlivosti použité při výpočtu intervalů spolehlivosti a hodnoty testovacích statistik vypočítaných při testech významnosti neodpovídají skutečnosti. V praxi tedy vychýlené hodnoty směrodatných chyb vedou k nesprávným úsudkům při používání dvou uvedených nástrojů frekvenčního přístupu ke statistické inferenci.

Jsou naše výsledky ve shodě s teoretickými očekáváními? V tomto případě se přirozeně nejedná o vlastní výzkumnou otázku. Odpověď na tuto otázku slouží čistě jako indikátor správnosti celého realizovaného postupu. Konkrétně tím máme na mysli, že v hlavních tendencích by měly být naše výsledky shodné s teoretickými očekáváními.

Například se zvyšujícím podílem chybějících hodnot by se při mechanismu NMAR také mělo zvyšovat vychýlení odhadovaných parametrů (Graham 2012: 239). Pokud by se naše výsledky lišily, tak je potřeba podrobně analyzovat potenciální příčiny těchto rozdílů.

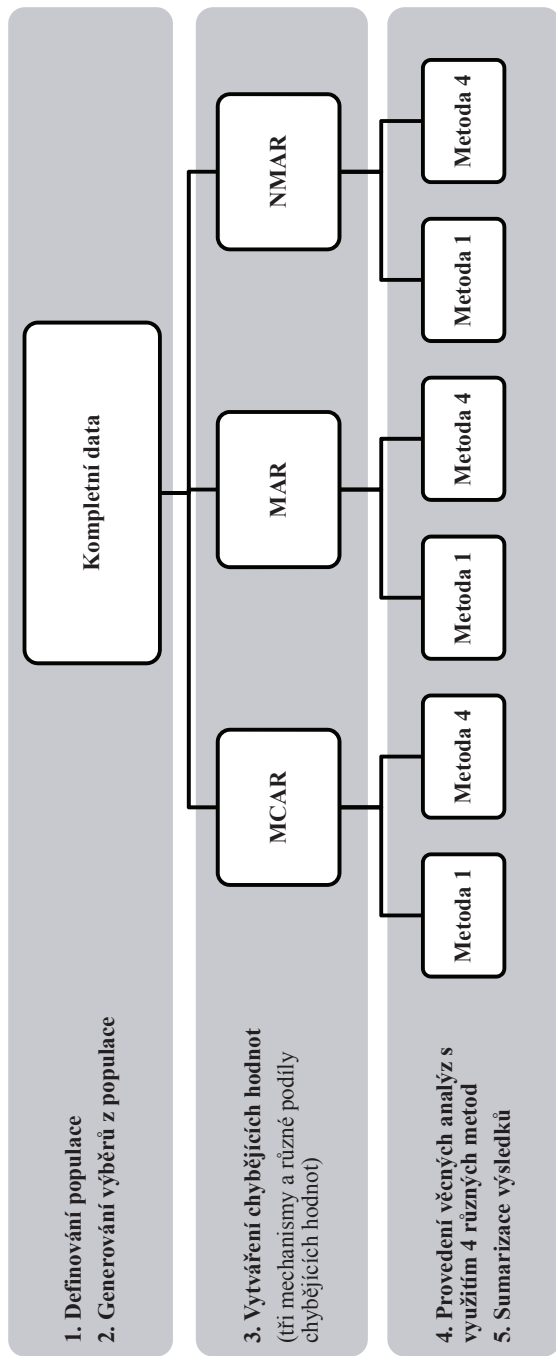
4.3 Postup realizace simulační studie

Obě simulační studie jsou podrobně popsány v příslušných kapitolách. Na tomto místě věnujeme prostor obecnému postupu, který je pro obě studie stejný. Na stanovené cíle analytické části knihy přímo navazuje určení experimentálních podmínek, tzn. okolností, které v rámci konkrétní simulace variiují, a naopak i podmínek, které jsou v rámci simulace konstantní¹⁷². Mezi často variované podmínky u obdobných simulačních studií patří rozsah výběru (např. Collins et al. 2001; Olinsky et al. 2003), počet proměnných zahrnutých do simulací a jejich pravděpodobnostní rozdělení (např. Little 1988, Demirtas et al. 2008), podíl chybějících hodnot (např. Olinsky et al. 2003, Demirtas et al. 2008), vzorec chybějících hodnot, mechanismy chybějících hodnot (např. Fox 2008: 554–555; Schafer a Graham 2002; Enders 2010: 52–54), metody práce s chybějícími hodnotami (např. Fox 2008: 554–555; Olinsky et al. 2003; Enders 2010: 52–54) a nakonec samotná věcná analýza a s ní související zkoumané parametry. U obou našich simulačních studií jsou experimentální podmínky zevrubně popsány zvlášť, přičemž nastavení některých podmínek se mezi studiemi liší.

Přejdeme teď k jednotlivým krokům technické realizace simulačních studií. První krok představuje definování populace. V tomto kroku je jednak určena věcně řešená analytická úloha a zároveň hodnoty populačních parametrů (tj. popisných charakteristik používaných proměnných a vztahů mezi těmito proměnnými). Po definování populace následuje generování výběrů (kompletních dat) z této populace. U obou simulačních studií se jedná o 10 000 výběrů z daných populací. V třetím kroku jsou z kompletních dat podle jednotlivých mechanismů chybějících hodnot vymazávány údaje. U třech mechanismů je přitom současně vytvářeno více podílů chybějících hodnot. Na takto vymazaných datech jsou potom realizovány věcné analýzy. Věcné analýzy jsou realizované vždy s použitím různých metod práce s chybějícími hodnotami. U obou simulačních studií se jedná vždy o čtyři metody práce s chybějícími hodnotami. Technicky i časově se jedná o nejnáročnější krok provedených simulací. Výsledky věcných analýz, které reprezentují aproximace výběrových rozdělení jednotlivých zkoumaných statistik, jsou uloženy do datových objektů (většinou se jedná o vektory) a následně sumarizovány. Schéma 4.1 přehledně shrnuje jednotlivé kroky technické realizace simulačních studií, tak jak byly uskutečněny v prostředí R.

¹⁷² Komplexnost realizovaných simulací vyjádřená počtem jedinečných kombinací experimentálních podmínek totiž exponenciálně stoupá. Proto v rámci simulačních studií variiují vždy jen některé podmínky a zbylé podmínky naopak mají konstantní hodnoty.

Schéma 4.1: Zjednodušený postup technické realizace simulační studie



Po provedení popsaných technických výpočtů nastává klíčová fáze interpretace výsledků. Při vyhodnocování výsledků je přirozeně kladený důraz na zodpovězení stanovených výzkumných otázek. Výsledky jsou vyhodnocovány na základě několika kritérií, kterým odpovídají v závěru této kapitoly představené indikátory. Obě simulační studie jsou vyhodnocovány s pomocí stejných nástrojů.

4.4 Statistický software a reprodukovatelnost simulací

Prezentované simulační studie byly realizované v prostředí pro statistické výpočty **R**. R je také programovací jazyk, který je prostřednictvím svobodné licence dostupný zdarma. Výhodou jazyka R je, že tento jazyk se používá jednak na interaktivní analýzu dat a zároveň i na programování (Chambers 2008: 13). Tato vlastnost byla využita i při tvorbě naší knihy.

Jednak byly totiž interaktivně používány existující funkce, které jsou součástí klíčových balíků¹⁷³, také však byly v rámci simulační studie vytvořené vlastní funkce, které umožnily snadnější řešení vzniklých analytických problémů. R dále umožňuje poměrně jednoduše vytvářet smyčky (loops)¹⁷⁴, v jejichž rámci mohou být používány různé datové objekty (např. vektory, matice či seznamy) a funkce. Jazyk R byl také zvolený kvůli poměrně jednoduché manipulaci s výsledky dílčích výpočtů a jejich přímočarému využití v dalších analýzách. V neposlední řadě představuje R díky své volné šířitelnosti ideální nástroj pro potenciální replikaci simulační studie. Na transparentnost a reprodukovatelnost realizovaných simulací byl přítom od začátku kladen velký důraz

Reprodukovatelnost je zajištěna jednak zpřístupněním celého skriptu v elektronické podobě¹⁷⁵, a zároveň kontrolováním generátoru pseudonáhodných čísel. Kontrolovat generátor pseudonáhodných čísel je v prostředí R možné s využitím dvou nástrojů (Chambers 2008: 227). Prvním je výběr metody generování pseudonáhodných čísel. R přitom disponuje možností výběru metody pro dva generátory, jeden pro uniformní rozdělení a druhý pro normální rozdělení (Chambers 2008: 227). V celé analytické části jsme pracovali s původními nastaveními obou metod generátorů náhodných čísel¹⁷⁶. Druhým nástrojem je tzv. nastavování seedu generátoru pseudonáhodných čísel (Chambers 2008: 228). Při každém využití generátoru pseudonáhodných čísel jsme

¹⁷³ Klíčové balíky neboli *core packages* jsou automaticky součástí prostředí a jsou k dispozici hned po spuštění programu. Jedná se o tyto klíčové balíky: *stats*, *graphics*, *grDevices*, *utils*, *datasets*, *methods*, *Autoloads* a *base*.

¹⁷⁴ V analytické části byla mnohokrát použita smyčka *for*, která realizovala požadované úkony vždy 10 000krát (což je počet opakování realizovaných simulací).

¹⁷⁵ Celý skript obsahuje více než 8000 řádků kódu (spolu s komentáři, které popisují jednotlivé kroky a tedy ulehčují orientaci v kódu). Kvůli uvedenému rozsahu není skript součástí tištěné verze knihy. Kompletní skript (včetně skriptu na tvorbu grafů) je dostupný na internetové adrese: <http://soc.cas.cz/publikace/analyza-chybejicich-hodnot>

¹⁷⁶ Konkrétně se jedná o metody "*Mersenne-Twister*" a "*Inversion*".

prostřednictvím funkce *set.seed* sami nastavili seed¹⁷⁷. Zároveň jsme při každém nastavení seedu generátoru použili jinou řadu přirozených čísel. Díky využití těchto dvou nástrojů je proto možné simulační studii kompletně zreprodukovat.

Reprodukovatelnost realizovaných studií zároveň umožňuje podrobit získané výsledky případným dalším analýzám. Nepřímo také poskytuje rámec pro variování dalších experimentálních podmínek, resp. k nastavení odlišných hodnot existujících experimentálních podmínek (např. podílů chybějících hodnot nebo dalších metod práce s chybějícími hodnotami).

4.5 Nástroje vyhodnocování výsledků simulací

Podívejme se tedy na jednotlivé nástroje, pomocí kterých vyhodnocujeme výsledky simulací. Konkrétně výsledky simulační studie vyhodnocujeme na základě následujících kritérií: průměrný odhad parametru, standardizované vychýlení, RMSE chyba, míra pokrytí populačního parametru intervaly spolehlivosti a průměrná délka intervalu spolehlivosti. Každý z těchto nástrojů je vždy zobrazený zvlášť v samostatné tabulce. Rozložení tabulek odpovídá variujícím experimentálním podmínkám. Tabulky mají analogickou strukturu, přičemž v samotných buňkách se mění zobrazovaný nástroj vyhodnocování.

Základní výsledek simulační studie představují průměrné odhady vybraných¹⁷⁸ populačních parametrů. Tyto odhady byly vypočítané pro všechny kombinace experimentálních podmínek. Jinými slovy, pro každou metodu práce s chybějícími hodnotami byl při jednotlivých mechanismech a podílech chybějících hodnot vypočítán průměr z odhadnutých parametrů v 10 000 výběrech. Tyto „základní popisné statistiky“ jsou zobrazeny v tabulkách 5.3, 5.4 a 6.3.

Pro posuzování vychýlení odhadovaných parametrů v různých experimentálních podmínkách používáme *standardizované vychýlení* (standardized bias). Všeobecně koncept vychýlení představuje odchylku realizovaného úsudku od skutečnosti (Everitt a Skrondal 2010: 45). V situaci určené našimi simulacemi reprezentuje vychýlení rozdíl mezi průměrným odhadem parametru a hodnotou parametru v populaci. Aby metoda práce s chybějícími hodnotami vedla k nevychýleným odhadům populačních parametrů, tak tento rozdíl musí být logicky malý. Při 10 000 opakováních simulací však není vůbec složité představit si, že i minimální rozdíl může být považován za statisticky významný (a tedy indikující vychýlení), i když o praktické významnosti nemůže být

¹⁷⁷ Pokud není seed nastavený uživatelem, tak R při inicializaci generátoru pseudonáhodných čísel využívá aktuální čas (Chambers 2008: 228). Takovéto ad hoc definování seedu proto zneumožňuje reprodukovatelnost analýzy.

¹⁷⁸ Schválně uvádíme, že se jedná o vybrané populační parametry. Jinými slovy, v rámci simulačních studií jsme neodhadovali všechny možné populační parametry (použité při definování populace), ale jenom ty parametry, které byly součástí věcně smysluplné analýzy (tzn. analýzy, která by byla pro výzkumníka věcně zajímavá).

řeč¹⁷⁹. Collins a spoluautoři uvádějí, že vliv vychýlení na intervaly spolehlivosti a testovací statistiky závisí na celkové rozptýlenosti odhadů (2001: 340). Rozptýlenost odhadů charakterizuje směrodatná chyba (směrodatná odchylka daného aproximovaného výběrového rozdělení). Collins a spoluautoři využívají směrodatnou chybu na standardizování vychýlení. Konkrétně tedy v souladu s těmito autory počítáme standardizované vychýlení podle vzorce (Collins et al. 2001: 340):

$$100 \frac{\text{průměrný odhad parametru} - \text{populační hodnota parametru}}{\text{směrodatná odchylka odhadů}}$$

Standardizované vychýlení proto vyjadřuje vzdálenost průměrného odhadu od populační hodnoty parametru jako procentuální podíl směrodatné chyby/směrodatné odchylky odhadů. Například standardizované vychýlení s hodnotou 100 vyjadřuje, že odhad parametru je v průměru o jednu směrodatnou chybu nad populační hodnotou parametru. To znamená, že se nachází o víc než čtvrtinu délky 95% intervalu spolehlivosti nad populační hodnotou parametru.

Collins a kolegové zároveň uvádějí, že pokud absolutní hodnota standardizovaného vychýlení překročí 40 % až 50 %¹⁸⁰, tak vychýlení začíná mít nepříznivý vliv hned na několik aspektů statistického usuzování (2001: 340). Vzhledem ke skutečnosti, že v našich simulacích pracujeme s větším rozsahem výběru¹⁸¹, tak jsme jako kritickou hranici pro standardizované vychýlení zvolili horní, 50% hranici¹⁸². S rostoucím rozsahem výběru se totiž ceteris paribus snižuje směrodatná odchylka odhadů, a tudíž při stejném absolutním vychýlení se zvyšuje standardizované vychýlení. Tabulky 5.5 a 6.4 zobrazují standardizované vychýlení vybraných odhadovaných parametrů při různých experimentálních podmínkách. Všechny buňky s hodnotou standardizovaného vychýlení nižší než -50 anebo vyšší než 50 jsou vybarvené šedou barvou. Jinými slovy, v těchto šedých buňkách je vychýlení prakticky významné.

Dalším často používaným kritériem při vyhodnocování úspěšnosti různých metod práce s chybějícími hodnotami je tzv. *RMSE chyba* (root mean square error)¹⁸³. Zjednodušeně řečeno, RMSE chyba představuje průměrnou odchylku odhadnutého parametru

¹⁷⁹ Podívejme se například na průměrný odhad průměru příjmu ($\mu_{\text{příjem}}$) při mechanismu MCAR, 30 % chybějících hodnot a nahrazováním chybějících hodnot stochastickou regresí. Po nahrazení chybějících hodnot stochastickou regresí se průměrný odhad rovná 24 991,6 Kč (údaj je z tabulky 5.3). Rozdíl mezi průměrným odhadem a populačním parametrem je jen 8,4 Kč. Představme si situaci, kdy počítáme na základě našich 10 000 výběrů, které aproximují výběrové rozdělení průměrného příjmu, jednovýběrový t-test s nulovou hypotézou $\mu_{\text{příjem}} = 25 000$ Kč. Při 9999 stupních volnosti odpovídá hodnotě testovací statistiky $t = -1,7216$ p-hodnota 0,0852 (jedná se o dvojitý test). To znamená, že jsme na hladině $\alpha = 0,1$ získali statisticky významný rozdíl. Mohli bychom tedy konstatovat, že vychýlení je v tomto případě statisticky významné (na hladině 0,1). O praktické významnosti rozdílu však při rozdílu jen 8,4 Kč není správné uvažovat.

¹⁸⁰ Stejné kritérium je možné najít i u jiných autorů (např. Demirtas et al. 2008).

¹⁸¹ Collins a spoluautoři pracují s 500 případy ($n = 500$) (2001: 339).

¹⁸² S polovinou směrodatné chyby odhadů jako kritériem pro posuzování problematičnosti vychýlení je možné se setkat i u Schafera s Grahamem (2002: 157).

¹⁸³ Její použití je možné nalézt například u Foxe (2008: 555), Schafera s Grahamem (2002) či Collinse a spoluautorů (2001).

tru od jeho populační hodnoty. Tato chyba se pro parametr θ a metodu práce s chybějícími hodnotami q vypočítá podle vzorce¹⁸⁴:

$$RMSE(\theta, q) = \sqrt{\frac{1}{N} \sum_{v=1}^N (\theta - \theta_{vq})^2},$$

kde θ_{vq} představuje odhad parametru θ metodou q ve výběru v . V každém výběru se tedy nejprve spočítá rozdíl mezi hodnotou populačního parametru θ a odhadem parametru v tomto výběru θ_{vq} . Po umocnění těchto rozdílů se z nich vypočítá průměrná hodnota. Po následném odmocnění tohoto průměru jsou hodnoty chyby RMSE ve stejných jednotkách jako hodnoty zkoumaného parametru. RMSE chyba nabývá jen kladných hodnot. Pokud metoda práce s chybějícími hodnotami vede k nevychýleným odhadům parametrů, tak se RMSE chyba rovná směrodatné chybě odhadů parametrů. Při vychýlených odhadech parametrů se RMSE chyba rovná součtu vychýlení a směrodatné chyby odhadů parametrů (Everitt a Skrondal 2010: 274). RMSE chyba proto v sobě kombinuje kromě konceptu vychýlení také koncept eficiency bodových odhadů¹⁸⁵ (Collins et al. 2001: 340). Pokud totiž dvě metody práce s chybějícími hodnotami současně poskytují nevychýlené odhady parametrů, tak metoda s nižší chybou RMSE má vyšší eficiency (protože rozptyl odhadů parametrů je u této metody menší). Hodnoty RMSE u vybraných parametrů při různých experimentálních podmínkách zobrazují tabulky 5.6 a 6.5.

Po vyhodnocení vychýlení (a eficiency) odhadů zaměříme svoji pozornost i na směrodatné chyby. Při zkoumání směrodatných chyb odhadů parametrů spojených s různými metodami práce s chybějícími hodnotami se tradičně používá indikátor procentuálního *pokrytí populačního parametru intervaly spolehlivosti* (např. Schafer a Graham 2002; Enders 2010: 96–97; Fox 2008: 555). Základní idea stojící za tímto indikátorem je následující: když daná metoda poskytuje nevychýlené odhady populačních parametrů, tak intervaly spolehlivosti by při dostatečném počtu opakování měly na zvolené hladině spolehlivosti pokrýt hodnotu populačního parametru právě v takovém podílu náhodných výběrů, který odpovídá zvolené hladině spolehlivosti. Například při počítání 95% intervalů spolehlivosti by měly spočítané intervaly spolehlivosti pokrýt populační hodnotu parametru právě v přibližně 95 % realizovaných výběrů. S ohledem na všeobecný vzorec výpočtu intervalů spolehlivosti proto takovéto pokrytí odpovídá nevychýleným odhadům směrodatných chyb parametrů (Enders 2010: 96).

Při použití 95% intervalů spolehlivosti je problematická míra pokrytí, která je nižší než 90 % (Schafer a Graham 2002: 157). Pokud se totiž při použití 95% intervalů spolehlivosti rovná skutečná míra pokrytí 90 %, tak se to projeví na nesprávné úrovni chyby prvního druhu. V takovéto situaci je pravděpodobnost chyby prvního druhu při tradičně zvolené hladině 0,05 ve skutečnosti dvojnásobná (Collins et al. 2001: 340;

¹⁸⁴ Tento vzorec je převzatý z dokumentace balíku *hydroGOF*, který byl použitý při výpočtu RMSE chyb (Zambrano-Bigiarini 2014: 60).

¹⁸⁵ Například Fox používá RMSE primárně jako indikátor eficiency (2008: 555).

Schafer a Graham 2002: 157). V rámci vyhodnocování naší simulační studie proto také u všech kombinací experimentálních podmínek počítáme 95% intervaly spolehlivosti. Svoji pozornost jsme však omezili jen na některé ze zkoumaných parametrů: průměrný příjem ($\mu_{\text{příjem}}$) a regresní koeficient (β) u první simulační studie a regresní koeficienty z modelu determinant politické znalosti u druhé simulační studie. Tabulky 5.7 a 6.6 zobrazují 95% intervaly spolehlivosti těchto parametrů. Míry pokrytí populačních parametrů vypočítané na kompletních datech odpovídají zvolené hladině spolehlivosti¹⁸⁶. V tabulkách 5.7 a 6.6 jsou kvůli jednodušší orientaci šedou barvou zvýrazněné buňky, kterým odpovídá míra pokrytí nižší než 90 %. Zároveň jsou tučným písmem zvýrazněné buňky, ve kterých daná metoda sice poskytuje nevychýlené odhady populačních parametrů, ale míra pokrytí je nižší než 90 %.

Poslední kritérium vyhodnocování výsledků simulací představuje *průměrná délka intervalu spolehlivosti*. Při vybraných parametrech byla na základě každého z 10 000 výběrů spočítána délka 95% intervalu spolehlivosti. Z těchto 10 000 délek se následně vypočítal průměr. Kratší průměrná délka přitom všeobecně odpovídá vyšší přesnosti daných intervalových odhadů. Proto jsou při dodržení předpokladu správné míry pokrytí populačního parametru preferovány kratší intervaly spolehlivosti (Schafer a Graham 2002: 157). Průměrné délky intervalů spolehlivosti jsou spolu s mírami pokrytí zobrazeny v tabulkách 5.7 a 6.6.

¹⁸⁶ K největšímu rozdílu mezi zvolenou 95% hladinou spolehlivosti a mírou pokrytí populačního parametru vypočítanou na kompletních datech dochází u regresního koeficientu pro pohlaví (druhá simulační studie). 95% intervaly spolehlivosti v této situaci pokryly hodnotu populačního parametru v 94,54 % výběrů.